Comparison of Classifiers for Classification of Arrhythmia

Stat610, Fall 2016, Robert Kramer

Data Set

UCI Arrhythmia Data Set [1]

- 279 Attributes (p)
  - 206 linear
  - 73 nominal
    - {0, 1} in data
- 452 Instances (n)
- includes missing values and zero mean / near zero mean features
- Dependent Variable
  - presence of cardiac arrhythmia
    - collapsed 15 types of arrhythmia into presence of arrhythmia

Description: [2]

Features were extracted from clinical records and statistics of standard 12 lead ECG recordings. Statistics include QRS duration, RR, P-R, and Q-T intervals.

The statistics of lead ECG recordings include features for the average QRS duration in msec.; the average duration between onset of P and Q waves in msec.; the average duration between onset of Q and offset of T waves in msec.; the average duration between two consecutive T waves in msec.; the average duration between two consecutive P waves in msec. Similar statistics were also chosen for each individual channel.

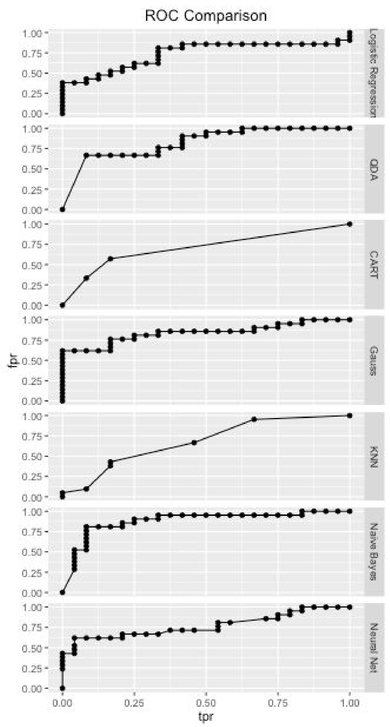
The classification problem of determining which type of arrhythmia is extremely unbalanced, but for simply determining the presence of arrhythmia there is an approximately even split.

Pre-Processing, Optimization, and Models

- Caret algorithm nearZeroMean was used to determine which features meet the conditions:
  1. $UniqueValues(Features) << n$
  2. $freq(mode)/freq(mode_{n-1}) -> large$
     - These features were eliminated
- Missing values were imputed using knnImpute function
  - Data is not in order and no reason to believe imputation will give an unbiased estimate of missing values.
  - Performed Imputation because n/p is small
- Data was centered and the mean set to zero
  - Wanted to have same preprocessing for all models
    - Both KNN and Naive Bayes are sensitive
  - Imputation technique applies standardization regardless
- All model parameters were optimized using 10 fold cross validation
  - Used default parameter tunegrid for each model
  - Preprocessing occurred inside CV loop
- Models Explored:
  - L2 Penalized Logistic Regression
  - QDA
  - CART
  - Gaussian Process
  - KNN
  - Naive Bayes
  - Additional Method - Neural Net (Multilayer Perceptron)

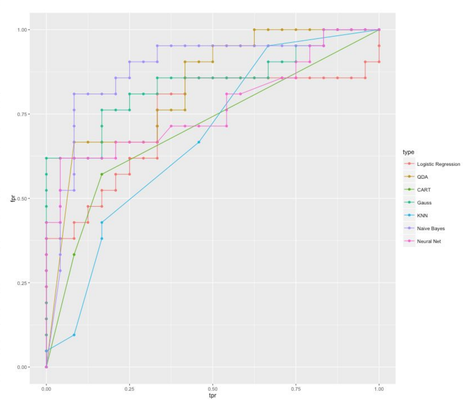
Results

Model	Train Error	Test Error	AUC	Accuracy	P-Value [ACC > NIR]
Logistic Regression	0.105651106	0.3333333	0.7361111	0.6667	0.049
QDA	0.007371007	0.3333333	0.8273810	0.6667	0.049
CART	0.257985258	0.2888889	0.7063492	0.7111	0.01151
Gaussian Process	0.100737101	0.2444444	0.8412698	0.7556	0.001836
KNN	0.260442260	0.3555556	0.6795635	0.6444	0.08860
Naive Bayes	0.255528256	0.2666667	0.8829365	0.7333	0.004840
Neural Net	0.046683047	0.3111111	0.7698413	0.6889	0.02486



Comparison:

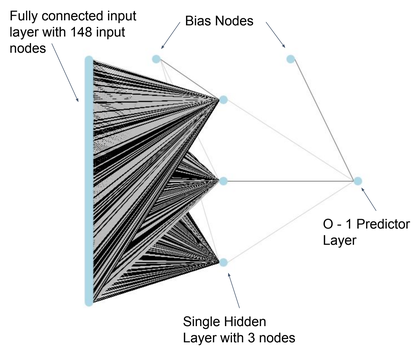
- Logistic Regression and QDA were equivalent on the test set
- Both CART and KNN gave similar results.
- The Neural Net clearly Overfit
- Our best overall classification method was the Naive Bayes according to AUC.
- Highest Accuracy was achieved by the Gaussian Process with a polynomial kernel



Multilayer Perceptron

- Equivalent to logistic regression in one layer
- The Neural Net model "nnet" chose has 3 hidden nodes and the lambda penalty parameter was set to zero.
  - Overfit; tried to force parameters, but results were poor
  - There were not enough training examples for the model to use more than 4 hidden nodes.

Plot of Fitted Model



Model: [6]

$$p(y|\mathbf{x},\boldsymbol{\theta}) = \text{Ber}(y|\text{sigm}(\mathbf{w}^T\mathbf{z}(\mathbf{x})))$$

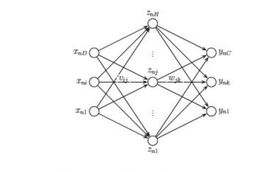


Figure 16.11 A neural network with one hidden layer.

Gradient Descent Backpropagation:

$$\delta_{nj}^v = \sum_{k=1}^{K} \delta_{nk}^w w_{kj} g'(a_{nj})$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_n [\boldsymbol{\delta}_n^v \mathbf{x}_n,\ \boldsymbol{\delta}_n^w \mathbf{z}_n]$$

Citations

1. UCI Machine Learning Repository: Arrhythmia Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Arrhythmia. (Accessed: 8th December 2016)
2. Guvenir, H. A., Acar, B., Demiroz, G. & Cekin, A. A supervised machine learning algorithm for arrhythmia analysis. in *Computers in Cardiology 1997* 433–436 (ieeexplore.ieee.org, 1997).
3. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28,** 1–26 (2008).
4. Prati, R. C., Batista, G. & Monard, M. C. Evaluating classifiers using ROC curves. *IEEE Latin America Transactions* **2,** 215–222 (2008).
5. Dietterich, T. G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **10,** 1895–1923 (1998).
6. Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT, 2012. Print.

Contact Information

kramer7@pdx.edu